

H-relative error estimation approach for multiplicative regression model with random effect

Zhanfeng Wang^{a,*}, Zhuojian Chen^a, Zimu Chen^a

^a*Department of Statistics and Finance, Management School, University of Science and Technology of China, Hefei, China.*

Abstract

Relative error approaches are more of concern compared to absolute error ones such as the least square and least absolute deviation, when it needs scale invariant of output variable, for example with analyzing stock and survival data. An h-relative error estimation method via the h-likelihood is developed to avoid heavy and intractable integration for a multiplicative regression model with random effect. Statistical properties of the parameters and random effect in the model are studied. To estimate the parameters, we propose an h-relative error computation procedure. Numerical studies including simulation and real examples show the proposed method performs well.

Keywords: Relative errors, Random effect, H-likelihood, asymptotic property

1. Introduction

In regression analysis, the least squares (LS) and least absolute deviation (LAD) are the most commonly used criteria based on absolute errors (Stigler, 1981; Portnoy et al., 1997). However, relative error methods are more of concern when it needs scale invariant of response variable (output) such as analyzing stock price and survival data (Narula and Wellington, 1977; Makridakis, 1985; Khoshgoftaar et al., 1992; Ye, 2007; Park and Stefanski, 1998; Chen et al., 2010; Zhang and Wang, 2013; Wang et al., 2015; Liu et al.,

*Department of Statistics and Finance, Management School, University of Science and Technology of China, Hefei, China. (Email: zfw@ustc.edu.cn).

2016; Chen et al., 2016). For example, based on multiplicative regression models, Chen et al. (2010) proposed a least absolute relative error (LARE) method, and Wang et al. (2015) developed a relative error change-point estimation approach. The LARE criterion was used to construct a local least absolute relative error estimation for a partially linear multiplicative model (Zhang and Wang, 2013). For a more flexible model, multiplicative single index model, Wang et al. (2016) showed a two-step estimation procedure to estimate the parameter and unknown link function with respect to relative errors.

However, the preceding relative error methods do not take a random effect account in their studied models. To the best of our knowledge, most of random effect approaches in literature are built on the absolute error or likelihood methods. It is much desired to study a relative error method for random effect models. When relative errors are of concern, the response variable generally is positive. Similar to the multiplicative regression model in Chen et al. (2016), we construct an h-relative error approach based on the following model

$$Y = \exp(X^T\beta + \nu)\epsilon, \quad (1)$$

where Y is the response variable, X is the p -vector of explanatory variables with the first component being 1 (intercept), β is the corresponding p -vector of regression parameters with the first component being the intercept, ν is the random effect and ϵ is the error term which is strictly positive. In model (1), when $\nu = 0$, Chen et al. (2016) proposed a least product relative error criterion which possesses some merits: smooth, convex and so on. They stated that under the error ϵ with the density

$$f(t) = c \exp\{-t - 1/t - \log(t) + 2\}I(t > 0), \quad (2)$$

the parameter estimator is asymptotically efficient, where c is a normalizing constant.

The density (2) is employed to develop a computation algorithm for the parameter estimation in this paper. The likelihood principle (Birnbaum, 1962) states that marginal likelihood of β carries all the information in the data about the fixed parameters β , so that the marginal likelihood should be used for inferences about β . However, in general the marginal likelihood requires intractable integration which is usually hard to obtain a precise result. One method to obtain the marginal maximum likelihood (ML) estimator for β is the expectation-maximization (EM) algorithm in Dempster et al. (1977).

But, the EM algorithm is often numerically slow to converge and other simulation methods, such as Monte Carlo EM (Vaida and Meng, 2004) and Gibbs sampling (Karim and Zeger, 1992) are computationally intensive. Instead, numerical integration using Gauss-Hermite quadrature (Crouch and Spiegelman, 1990) could be directly applied to obtain the ML estimators, but this also becomes computationally heavier as number of random components increases.

To avoid heavy and intractable integration, in this paper, we employ the h-likelihood method (Lee and Nelder, 1996, 2001, 2005) to build an h-relative error method to estimate the parameter β and the random effect ν . The proposed method also inherits scale invariance and less sensitive to outliers (Chen et al., 2010). We develop asymptotic properties of β and ν , such as consistence and normality. A computation algorithm is proposed to estimate the parameters via the h-relative errors. Numerical studies show the proposed method has better performance of the parameter estimation compared to the traditional linear mixed model.

The rest of this paper is organized as follows. Section 2 introduces relative errors, parameter estimation for model (1), and provides their statistical properties. Numeric studies including simulation and real examples are in Section 3. All of proofs of the theorems are presented in Appendix.

2. Methodologies

2.1. H-likelihood with relative errors

Throughout this paper we mean by c the positive constant independent of n , which may take different values in different formulae or even in different parts of one and the same inequality.

Suppose observation samples (Y_{ij}, X_{ij}) , $i = 1, \dots, K, j = 1, \dots, n_i$ are randomly generated from model (1) with repeatedly measured responses

$$Y_{ij} = \exp(X_{ij}^\top \beta + \nu_i) \epsilon_{ij}, \quad i = 1, \dots, K, j = 1, \dots, n_i. \quad (3)$$

It follows that $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$ is the response for the i th unit ($i = 1, \dots, K$) and ν_i is the corresponding unobserved random effect. Let the total sample size $n = \sum_{i=1}^K n_i$. Without loss of generality, let $\nu_i \sim N(0, \sigma^2)$ in this paper. The proposed methods can be extended to models with random effect having other distributions. Since ϵ_{ij} has the density from function (2), conditional density function $f(Y_{ij}|\nu_i; \theta)$ satisfies

$$\log(f(Y_{ij}|\nu_i; \beta)) \equiv -Y_{ij} \exp(-X_{ij}^\top \beta - \nu_i) - \exp(X_{ij}^\top \beta + \nu_i)/Y_{ij} + c,$$

which suggests that

$$-\log(f(Y_{ij}|\nu_i; \beta)) = \left| \frac{Y_{ij} - \exp(X_{ij}^\top \beta + \nu_i)}{Y_{ij}} \right| \times \left| \frac{Y_{ij} - \exp(X_{ij}^\top \beta + \nu_i)}{\exp(X_{ij}^\top \beta + \nu_i)} \right| + c. \quad (4)$$

It shows that

$$\left| \frac{Y_{ij} - \exp(X_{ij}^\top \beta + \nu_i)}{Y_{ij}} \right| \quad \text{and} \quad \left| \frac{Y_{ij} - \exp(X_{ij}^\top \beta + \nu_i)}{\exp(X_{ij}^\top \beta + \nu_i)} \right|$$

are two types of relative error: one is the error relative to the target and the other is the error relative to the predictor of the target. It shows that (4) consists of these relative errors, which leads to a relative error estimation criterion with respect to likelihood technique.

For convenience of notations, let $\theta = (\beta^\top, \sigma^2)^\top$, $\nu = (\nu_1, \nu_2, \dots, \nu_K)^\top$ and $Y = (Y_1^\top, Y_2^\top, \dots, Y_K^\top)^\top$, where β denotes the location parameter and σ^2 denotes the dispersion parameter. For model (3), a log h-likelihood on the parameters is defined as

$$H\{\theta, \nu; Y\} = \sum_{i=1}^K h_i\{\theta, \nu_i; Y_i\} = \sum_{i=1}^K \left(l_{1i}\{\theta, \nu_i; Y_i\} + l_{2i}\{\theta, \nu_i\} \right), \quad (5)$$

where

$$l_{1i}\{\theta, \nu_i; Y_i\} = \sum_{j=1}^{n_i} \log f(Y_{ij}|\nu_i; \beta),$$

$$l_{2i}\{\theta, \nu_i\} = \log f(\nu_i; \sigma^2) = -\frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \nu_i^2.$$

It easily shows that $\sum_{i=1}^K l_{1i}\{\theta, \nu_i; Y_i\}$ becomes the least product relative error criterion (LPRE) in Chen et al. (2016).

To introduce the connection between data generation and parameter estimation, this paper also considers the extended likelihood framework (Lee and Nelder, 2005)

$$L(\theta, \nu; y, \nu) \equiv f_\theta(\nu, y) = f_\theta(\nu) f_\theta(y|\nu) = f_\theta(y) f_\theta(\nu|y).$$

Then $H\{\theta, \nu; Y\}$ can be rewritten as

$$H\{\theta, \nu; Y\} = m(\theta, Y) + l(\theta, \nu|y), \quad (6)$$

where

$$m(\theta, Y) = \log L(\theta, Y) = \sum_{i=1}^K \log \int e^{h_i\{\theta, \nu_i; Y_i\}} d\nu_i = \sum_{i=1}^K m_i(\theta, Y), \quad (7)$$

$$l(\theta, \nu|y) = \sum_{i=1}^K l_i(\theta, \nu_i|y) = \sum_{i=1}^K \log \frac{f(Y_i|\nu_i; \beta)f(\nu_i; \sigma^2)}{\int e^{h_i(\theta, \nu_i; Y_i)} d\nu_i}. \quad (8)$$

It follows that $m(\theta, Y)$ is a marginal log-likelihood with respect to θ and $l(\theta, \nu|y)$ is a conditional density function of (θ, ν) for given data Y .

Estimators of the random effect ν_i and parameter θ are obtained by maximizing the log h-likelihood (5) or the log extended likelihood (6), that is why we call h-relative error estimation approach for multiplicative regression model with random effect.

2.2. Inference on random effects

We firstly treat θ as known. Inference on ν_i with an estimate of θ will be discussed later in this section. We know estimation of ν_i only involves information from the i th subject when θ is fixed. From the h-likelihood method (Lee and Nelder, 1996), it gives an estimator of ν_i , saying $\hat{\nu}_i$, by solving the equation

$$h_i^{(1)}\{\theta, \nu_i; Y_i\} = 0,$$

where $h_i^{(k)}\{\theta, \nu_i; Y_i\} = \partial^k h_i / \partial \nu_i^k$, $k = 1, 2, \dots, 6$.

For $r = r(\nu)$, the quantity $\delta = E(r|y)$ is the best unbiased predictor for the r in the sense that $E(\delta) = E_y E(r|y) = E(r)$. And it has the minimum mean-square error of prediction with respect to $E(\delta - r)'P(\delta - r)$ for any positive define matrix P .

Under appropriate conditions, we show that $\hat{\nu}$ converges to $E(\nu|y)$ presented in the following theorem, which proof is in Appendix.

Theorem 1. *Under conditions A_1 in Appendix hold, we have*

$$\hat{\nu}_i = E(\nu_i|Y) + O_p\left(\frac{1}{n}\right), \quad \text{Var}(\nu_i|Y) = D_i^{*-1}\{1 + O_p(n^{-1})\},$$

where $D_i^* = -\partial^2 H / \partial \nu_i^2|_{\nu_i=\hat{\nu}_i}$.

For given θ , the Laplace approximations of the expressions (5) and (6) with respect to the random effect are

$$\hat{H} \propto h_i\{\theta, \hat{\nu}_i; Y_i\} - \frac{1}{2}(\hat{\nu}_i - \nu_i)' D_i^* (\hat{\nu}_i - \nu_i), \quad (9)$$

$$\hat{H} \propto m_i(\theta; Y_i) + \hat{l}_i(\theta; \nu_i | Y_i), \quad (10)$$

where \hat{H} and \hat{l}_i are separately Taylor expansions of H and l_i with respect to ν_i at point $\nu_i = \hat{\nu}_i$. Since m_i and $h_i\{\theta, \hat{\nu}_i; Y_i\}$ do not depend on ν_i , they can be ignored when the distribution of $\nu_i | Y$ is computed. Thence, (9) and (10) imply that

$$\nu_i | Y \propto N(\hat{\nu}_i, D_i^{*-1}).$$

It follows that $N(\hat{\nu}_i, D_i^{*-1})$ is a reasonable approximation distribution of $\nu_i | Y$. Easily we show that D_i^{*-1} has order of $O_p(n_i^{-1})$ under the assumption A_1 .

Under unknown θ , following Paik et al. (2015), let $(\hat{\theta}, \hat{\nu})$ be a solution of

$$\left(\frac{\partial}{\partial \theta} m(\theta; Y) \right)_{W\{\theta, \nu; Y\}} = 0, \quad (11)$$

where $W\{\theta, \nu; Y\} = (h_1^{(1)}\{\theta, \nu_1; Y_1\}, \dots, h_K^{(1)}\{\theta, \nu_K; Y_K\})^\top$. For a realized value ν_{0i} , we can have the next theorem, which is similar to Paik et al. (2015).

Theorem 2. *Under Conditions A_1 and A_2 holds, $\sqrt{n_i}(\hat{\nu}_i - \nu_{0i})$ converges in distribution to normal with mean 0 and variance*

$$I(\theta, \nu_{0i})^{-1} + n_i I(\theta, \nu_{0i})^{-1} B_{21i} A_{11}^{-1} \text{Var} \left[\frac{\partial}{\partial \theta} m_i(\theta; Y_i) | \nu_i = \nu_{0i} \right] A_{11}^{-1} B_{21i}^\top I(\theta, \nu_{0i})^{-1} \\ - 2n_i I(\theta, \nu_{0i})^{-1} B_{21i}^\top A_{11}^{-1} \text{Cov} \left[\frac{\partial}{\partial \theta} m_i(\theta; Y_i) | \nu_i, h_i^{(1)}\{\theta, \nu_i; Y_i\} | \nu_i = \nu_{0i} \right],$$

where $A_{11} = E\{-\frac{\partial^2}{\partial \theta \partial \theta^\top} m_i(\theta; Y_i)\}$, $B_{21i} = \frac{1}{n_i} E\{\frac{\partial}{\partial \theta} h_i^{(1)}\{\theta, \nu_i; Y_i\} | \nu_i = \nu_{0i}\}$ and $I(\theta, \nu_{0i}) = \frac{1}{n_i^2} E\left[-h_i^{(2)}\{\theta, \nu_i; Y_i\} | \nu_i = \nu_{0i}\right]$.

2.3. Inference on location parameters

Naturally, one way to estimate the parameter β is to maximize the marginal log-likelihood $m(\theta, Y)$, but the integration involved in $m(\theta, Y)$ is intractable. Following Paik et al. (2015), we use a Laplace approximation

method to compute the integration. For convenience of notations, let the Laplace approximation of a function $l(\alpha)$ on α , be

$$p_\alpha(l) = \left[l - \frac{1}{2} \log \det \{ D(l, \alpha) / (2\pi) \} \right] \Big|_{\alpha=\tilde{\alpha}} \quad (12)$$

where $D(l, \alpha) = -\partial^2 l / \partial \alpha^2$ and $\tilde{\alpha}$ is one solution of the equation $\partial l / \partial \alpha = 0$.

Next we show that $p_\nu(H)$ is a reasonable approximation to $m(\theta, Y)$. From Tierney and Kadane (1986), we show that

$$\exp\{m_i(\theta; Y_i)\} = \int e^{h_i\{\theta, \nu_i; Y_i\}} d\nu_i = e^{h_i\{\theta, \hat{\nu}_i; Y_i\}} \sqrt{2\pi\tau_i n_i^{-1/2}} [1 - C_{n_i}\{\theta, \hat{\nu}_i\}] + O(n_i^{-2}), \quad (13)$$

where

$$\begin{aligned} \tau_i^2 &= -[h_i^{(2)}\{\theta, \hat{\nu}_i; Y_i\}]^{-1}, \\ C_{n_i}\{\theta, \hat{\nu}_i\} &= J_{1i}\{\theta, \hat{\nu}_i; Y_i\} / (8n_i) - 5J_{2i}\{\theta, \hat{\nu}_i; Y_i\} / (24n_i), \\ J_{1i}\{\theta, \hat{\nu}_i; Y_i\} &= -h_i^{(4)}\{\theta, \hat{\nu}_i; Y_i\} / [h_i^{(2)}\{\theta, \hat{\nu}_i; Y_i\}]^2, \\ J_{2i}\{\theta, \hat{\nu}_i; Y_i\} &= -[h_i^{(3)}\{\theta, \hat{\nu}_i; Y_i\}]^2 / [h_i^{(2)}\{\theta, \hat{\nu}_i; Y_i\}]^3. \end{aligned}$$

For model (3), it easily shows that C_{n_i} , $J_{1i}\{\theta, \hat{\nu}_i; Y_i\}$ and $J_{2i}\{\theta, \hat{\nu}_i; Y_i\}$ have the order $O_p(1/n_i)$. Therefore, we obtain

$$\begin{aligned} m(\theta; Y) &= \sum_{i=1}^K m_i(\theta; Y_i) = \sum_{i=1}^K [h_i\{\theta, \hat{\nu}_i; Y_i\} - \frac{1}{2} \sum_{i=1}^K \log[-h_i^{(2)}\{\theta, \hat{\nu}_i; Y_i\} / 2\pi] \\ &\quad + \sum_{i=1}^K \log[1 - C_{n_i}\{\theta, \hat{\nu}_i\}] + O_p(n^{-1}). \end{aligned} \quad (14)$$

The first two terms in (14) are also called the adjusted profile likelihood. The equation (14) indicates the following theorem,

Theorem 3. *Under Conditions in Appendix, the marginal likelihood*

$$m(\theta; Y) = p_\nu(H) + O_p\left(\frac{1}{n}\right).$$

From Theorem 3, $p_\nu(H)$ is a perfect approximation to the marginal log-likelihood $m(\theta; Y)$. Hence, we directly give an estimator of β by maximizing $p_\nu(H)$ with respect to β instead of $m(\theta; Y)$. Let $\tilde{\beta}$ and $\hat{\beta}$ be maximizers of $m(\theta; Y)$ and $p_\nu(H)$, respectively. Next theorem shows a connection between $\tilde{\beta}$ and $\hat{\beta}$, which proof is in Appendix.

Theorem 4. *Under Conditions in Appendix, for given ν and σ^2 , we have*

$$\tilde{\beta} = \hat{\beta} + O\left(\frac{1}{n}\right).$$

Then we can use $\hat{\beta}$ as an estimator of β . To make inference on β , we need to know the variance of $\hat{\beta}$. Let

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = n \begin{pmatrix} Var(\hat{\beta}) & Cov(\hat{\beta}, \hat{\nu} - \nu) \\ Cov(\hat{\nu} - \nu, \hat{\beta}) & Var(\hat{\nu}) \end{pmatrix},$$

$$M = \frac{1}{n} \begin{pmatrix} -\frac{\partial^2 H}{\partial \beta \partial \beta^T} & -\frac{\partial^2 H}{\partial \beta \partial \nu^T} \\ -\frac{\partial^2 H}{\partial \nu \partial \beta^T} & -\frac{\partial^2 H}{\partial \nu \partial \nu^T} \end{pmatrix} \bigg|_{\beta=\hat{\beta}, \nu=\hat{\nu}}.$$

If $E(M)$ is non-singular, under appropriate conditions we show that M^{-1} converges to V as $n \rightarrow \infty$. Thence, M^{-1} can be used to estimate the variance of $\hat{\beta}$. We use the result of Lee and Nelder (1996) and we can see the proof in their appendix.

2.4. Estimation of dispersion parameter

It is well-known that for mixed linear models, in order to reduce bias, a restricted log-likelihood (Patterson and Thompson, 1971) is used to estimate the dispersion parameters. For model (3), the restricted log-likelihood is

$$r = \log L(\sigma^2; Y|\hat{\beta}) \equiv \log f_{\sigma^2}(Y|\hat{\beta}).$$

For mixed linear model, Cox and Reid (1987) extended r to $p_{\beta}(m)$. To avoid intractable integration in $p_{\beta}(m)$, following Lee and Nelder (2001) we use $p_{\beta, \nu}(h)$ to approximate $p_{\beta}(m)$. Maximizing $p_{\beta, \nu}(h)$ gives an estimate of the dispersion parameter. We know that logarithm transformation of model (3) is a mixed linear model such that h-likelihood for the logarithm transformation model differs only a constant (in Jacobi matrix) from $H\{\theta, \nu; Y\}$. Hence, from Lee and Nelder (2001), maximizing $p_{\beta, \nu}(h)$ provides a reasonable dispersion estimators. For further details, please see Lee and Nelder (2001).

2.5. Inference procedure

From suggestion in Lee and Nelder (2005), we generally use the h-loglihood H , the marginal likelihood m and the restricted loglihood $p_{\beta}(m)$ for inference of ν , β and σ^2 , respectively. Traditionally, we use sampling method

like Monte Carlo simulation to calculate m . In our method, to avoid the calculation of integration m , we use $p_\nu(h)$ and $p_{\nu,\beta}(h)$ to estimate β and σ^2 instead of m and $p_\beta(m)$. Therefore, the estimation equations are

$$\begin{cases} \frac{\partial H}{\partial \nu} = 0 \\ \frac{\partial p_\nu(H)}{\partial \beta} = 0 \\ \frac{\partial p_{\nu,\beta}(H)}{\partial \sigma^2} = 0. \end{cases} \quad (15)$$

Iteration algorithm such as Newton-Raphson algorithm is applied to solve the equation (15). From our simulation results, the estimation procedure converges quickly, for example with two or three iterations. However, when number of repeated measurement K and sample size n are large, computation of the matrices involved in estimation procedure becomes much more complicated. Under this case, we can compile the subprogram to overcome this shortcoming by using C or python language.

2.6. A mixed linear model with known variance of error

Since the error term ϵ in model (3) have a specific distribution in (2), we obtain density function of $\log(\epsilon)$

$$f(t) = c \exp(-e^t - e^{-t} + 2), \quad (16)$$

where $c \approx 0.594$ is a normalizing constant. By Taylor expansion on e^t , it is amazing to find $\log(\epsilon)$ behaves almost like a normal distribution with mean 0 and standard variance $\phi = 0.6434$. So we also compare the proposed method with the following logarithm transformation model

$$\tilde{Y}_{ij} = \log(Y_{ij}) = X_{ij}^T \beta + \nu_i + e_{ij}, i = 1, \dots, K, j = 1, \dots, n, \quad (17)$$

where e_{ij} has normal distribution with mean 0 and standard variance ϕ . Let $\tilde{Y}_i = (\tilde{Y}_{i1}, \dots, \tilde{Y}_{in})^T$ and $\tilde{Y} = (\tilde{Y}_1^T, \dots, \tilde{Y}_K^T)^T$. It easily shows that $\Sigma = \text{Cov}(Y) = \sigma^2(I_K \otimes 1_n 1_n') + \phi^2 I_{Kn}$, where 1_n is a length n vector with all elements 1, I_q is an identical matrix with rank q , and \otimes stands for Kronecker product.

To fit model (17), the least square (LS) method is used to provide equation

$$X^T \Sigma^{-1} X \beta = X^T \Sigma^{-1} \tilde{Y}, \quad (18)$$

where $X = (X_1^T, \dots, X_K^T)^T, X_i = (X_{i1}, \dots, X_{in})^T$. Solving (18) gives an estimate of β .

3. Numerical study

3.1. Simulation study

Simulation studies are constructed to evaluate finite sample performance of the proposed method. Simulation data are generated independently from model (1), where covariates have uniform distribution on $(0, 1)$ and random effect have standard normal distribution. The error term has three distributions: E_1 , E_2 and E_3 , where E_1 is (2) , E_2 is an exponential normal distribution with mean 0 and variance 0.414, and E_3 is an exponential of the uniform distribution on $(-2, 2)$. Sample size $(K, n_i) = (10, 10), (20, 5)$ and $(20, 10)$. Parameter $\beta^\top = (\alpha, \beta_1, \beta_2, \beta_3)^\top = (2, 2, 1, 1)^\top$. All of simulation are repeated 500 times.

Following Paik et al. (2015), we separately consider cases of known and unknown dispersion parameter σ^2 , where σ^2 is settled with value 1 for the known case while needs to be estimated for the unknown one. Here, performance of the proposed h-likelihood relative error method (HRE) is compared with that of the traditional least square method (LSE) mentioned in subsection 2.6. Tables 1 and 2 present results of parameter estimation from HRE and LSE for the known and unknown dispersion parameters, respectively. We can see that all estimates of parameter β from HRE and LSE are very close to their true values. However, under E_3 , HRE has smaller standard deviation than LSE, while under E_1 and E_2 these two methods have comparable results. From Table 2, under E_3 estimates of σ^2 have larger bias for LSE while HRE performs well. As K or n_i increases, the standard deviations of the parameter estimates become smaller.

3.2. Application

The proposed method is applied to two datasets, cakes data and sleep-study data. The cake data contains $n_i = 6$ different baking temperatures ranged from $175^\circ C$ to $225^\circ C$, and three different recipes. Among each recipe, there were $K = 15$ replications. It assumes that these replications have a randomized blocks scheme: one by one is produced, so that the differences among replicates may represent time effect. The response here is breaking angle, covariate is temperature. Since the breaking angle is gradual, it tends to have a subjective element (random effect). We can find sleepstudy data in package lme4. The average reaction time per day for each subject is recorded in a sleep deprivation study. On day 0 the $K = 18$ subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep

per night and were measured $n_i = 10$ days. The responses are the average reaction time on a series of tests in each day, covariate is date and random effect is brought in by subjective effect.

To evaluate the performance of HRE and LSE, each dataset is partitioned into two parts: around 2/3 samples as training data and the left as test data. The prediction accuracies from these two methods are measured by four different median indices: median of absolute prediction errors $\{|Y_i - \hat{Y}_i|\}$ (MPE), median of product relative prediction errors $\{|Y_i - \hat{Y}_i|^2 / Y_i \hat{Y}_i\}$ (MPPE), median of additive relative prediction errors $\{|Y_i - \hat{Y}_i| / Y_i + |Y_i - \hat{Y}_i| / \hat{Y}_i\}$ (MAPE) and median of squared prediction errors $\{|Y_i - \hat{Y}_i|^2\}$ (MSPE), where $\hat{Y}_{ij} = \exp(X_{ij}^\top \hat{\beta} + \hat{\nu}_i)$ for HRE. Parameter estimates and prediction results are shown in Table 3. It shows that HRE and LSE have similar estimates for β , but HRE has much smaller variance estimate than LSE. For cakes data, the breaking angle is larger when temperature increases, and for sleepstudy data, the average reaction time with longer studied time is larger. More importantly, all these 4 prediction indices from HRE are smaller than LSE.

References

- Birnbaum, A., 1962. On the foundations of statistical inference. *Journal of the American Statistical Association* 57 (298), 269–306.
- Chen, K., Guo, S., Lin, Y., Ying, Z., 2010. Least Absolute Relative Error Estimation. *Journal of the American Statistical Association* 105 (491), 1104–1112.
- Chen, K., Lin, Y., Wang, Z., Ying, Z., 2016. Least product relative error estimation. *Journal of Multivariate Analysis* 144, 91–98.
- Cox, D. R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–39.
- Crouch, E. A., Spiegelman, D., 1990. The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(x) \exp(-t^2)$: Application to logistic-normal models. *Journal of the American Statistical Association* 85 (410), 464–469.
- Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

- Karim, M. R., Zeger, S. L., 1992. Generalized linear models with random effects; salamander mating revisited. *Biometrics*, 631–644.
- Khoshgoftaar, T. M., Bhattacharyya, B. B., Richardson, G. D., 1992. Predicting software errors, during development, using nonlinear regression models: a comparative study. *IEEE Transactions on Reliability* 41 (3), 390–395.
- Lee, Y., Nelder, J. A., 1996. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 619–678.
- Lee, Y., Nelder, J. A., 2001. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* 88 (4), 987–1006.
- Lee, Y., Nelder, J. A., 2005. Likelihood for random-effect models. *Sort: Statistics and Operations Research Transactions* 29 (2), 141–182.
- Liu, X., Lin, Y., Wang, Z., 2016. Group variable selection for relative error regression. *Journal of Statistical Planning and Inference* 175, 40–50.
- Makridakis, S. G., 1985. The forecasting accuracy of major time series methods. *Journal of the Royal Statistical Society. Series D (The Statistician)* 34 (2), 261–262.
- Narula, S. C., Wellington, J. F., 1977. Prediction, linear regression and the minimum sum of relative errors. *Technometrics* 19 (2), 185–190.
- Paik, M. C., Lee, Y., Ha, I. D., 2015. Frequentist inference on random effects based on summarizability. *Stat. Sinica* 25, 1107–1132.
- Park, H., Stefanski, L., 1998. Relative-error prediction. *Statistics & probability letters* 40 (3), 227–236.
- Patterson, H. D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* 58 (3), 545–554.
- Portnoy, S., Koenker, R., et al., 1997. The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science* 12 (4), 279–300.

- Stigler, S. M., 1981. Gauss and the invention of least squares. *The Annals of Statistics*, 465–474.
- Tierney, L., Kadane, J. B., 1986. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association* 81 (393), 82–86.
- Vaida, F., Meng, X., 2004. Mixed linear models and the em algorithm in applied bayesian and causal inference from an incomplete data perspective. John Wiley and Sons.
- Wang, Z., Chen, Z., Wu, Y., Dec. 2016. A relative error estimation approach for single index model. *ArXiv* <https://arxiv.org/abs/1609.01553>.
- Wang, Z., Liu, W., Lin, Y., 2015. A change-point problem in relative error-based regression. *TEST* 24 (4), 835–856.
- Ye, J., 2007. Price models and the value relevance of accounting information. Available at SSRN 1003067.
- Zhang, Q., Wang, Q., 2013. Local least absolute relative error estimating approach for partially linear multiplicative model. *Statistica Sinica*, 1091–1116.

Appendix: Proofs of the main results

Condition A_1 : $n_i \rightarrow \infty$, $n_i/K \rightarrow O_p(1)$.

Condition A_2 : $||\frac{\partial}{\partial \theta} W_i^{(1)}\{\theta, \nu_i; Y_i\}|| = O_p(1)$.

Proof of Theorem 1

The Laplace approximation defined in (12) each group i becomes

$$\hat{h}_i = \hat{m}_i(\theta; Y_i) + \hat{l}_i(\theta; \nu_i | Y_i) \quad (\text{A.1})$$

where h_i is the h-likelihood, m_i is the marginal likelihood of Y_i and $\hat{l}_i(\theta; \nu_i | Y_i)$ is the conditional likelihood of ν_i given Y_i . Using the power series expansion we can get

$$\exp h_i = \exp \hat{h}_i \{1 + c_3(\nu_i - \hat{\nu}_i)^3 + c_4(\nu_i - \hat{\nu}_i)^4 + \dots\} \quad (\text{A.2})$$

where $c_3 = \frac{1}{6} \frac{h_i^{(3)}(\hat{\nu}_i)}{h_i(\hat{\nu}_i)}$ and $c_4 = \frac{1}{24} \frac{h_i^{(4)}(\hat{\nu}_i)}{h_i(\hat{\nu}_i)}$. Because we have specific expression and $Y_{ij} > 0$, we can get that c_3 and c_4 are coefficients with $O_p(n_i)$ order, or even $O_p(1)$. Tierney and Kadane (1986) showed that $\exp m_i = \exp \hat{m}_i \{1 + O_p(n_i^{-1})\}$. So we can show that $l + O_p(n_i^{-1}) = \hat{l} \{1 + c_3(\nu_i - \hat{\nu}_i)^3 + c_4(\nu_i - \hat{\nu}_i)^4 + \dots\}$, therefore $l_i = \hat{l}_i \{1 + c_3(\nu_i - \hat{\nu}_i)^3 + c_4(\nu_i - \hat{\nu}_i)^4 + \dots O_p(n_i^{-1})\}$. Because $\hat{l}(\theta; \nu_i | Y_i)$ is the log-likelihood of the normal density, therefore $E(\nu_i | Y_i) = \hat{\nu}_i + (c_3 - \hat{\nu}_i c_4) E^*(\nu_i - \hat{\nu}_i)^4 = \hat{\nu}_i + O_p(n_i^{-1})$

Proof of Theorem 4

Using the fact that

$$\int f_\theta(\nu | y) d\nu = 1 \quad (\text{A.3})$$

we can get the conclusion that

$$E(\partial h / \partial \theta | y) = \partial m / \partial \theta + E(\partial \log f_\theta(\nu | y) / \partial \theta | y) = \partial m / \partial \theta \quad (\text{A.4})$$

Consider the Taylor series expansion

$$\partial h / \partial \beta_k = \partial h / \partial \beta_k|_{\nu=\hat{\nu}} + A_1(\nu - \hat{\nu}) + A_2(\nu - \hat{\nu})^2 / 2! + \dots, \quad (\text{A.5})$$

where $A_1 = (\partial / \partial \beta_k)(\partial h / \partial \nu)|_{\nu=\hat{\nu}}$ and $A_2 = (\partial / \partial \beta_k)(\partial^2 h / \partial \nu^2)|_{\nu=\hat{\nu}}$. Since $\hat{\nu} = E(\nu | y) + O_p(n^{-1})$, $\text{var}(\nu | y) = O_p(n^{-1})$ and $A_i = O_p(n)$, Equation B.3 becomes $E\{\partial h / \partial \beta_k | y\} = \partial h / \partial \beta_k|_{\nu=\hat{\nu}} + O_p(1)$. Let $\hat{\beta}_k$ be the solution of $\frac{\partial m}{\partial \beta_k} = 0$ and $\tilde{\beta}_k$ be the solution of $\frac{\partial p_\nu(h)}{\partial \beta_k} = 0$, Equation A.5 becomes

$$\frac{\partial m}{\partial \beta_k} = \frac{\partial h}{\partial \beta_k}|_{\nu=\hat{\nu}} + O_p(1) \quad (\text{A.6})$$

and we can use Taylor series expansion again and equation B.4 becomes

$$\begin{aligned} \frac{\partial m}{\partial \hat{\beta}_k} &= \frac{\partial h}{\partial \hat{\beta}_k}|_{\nu=\hat{\nu}} + O_p(1) \\ &= \frac{\partial h}{\partial \tilde{\beta}_k}|_{\nu=\hat{\nu}} + (\hat{\beta} - \tilde{\beta}) \frac{\partial^2 h}{\partial \tilde{\beta}_k^2}|_{\nu=\hat{\nu}} + \dots + O_p(1) \\ &= 0 \end{aligned}$$

Since $\frac{\partial^{(n)} h}{\partial \tilde{\beta}_k^{(n)}}|_{\nu=\hat{\nu}}$ are of $O_p(n)$ order, we can get the conclusion that $\hat{\beta} - \tilde{\beta}$ is of $O_p(\frac{1}{n})$ order.

Table 1: Results of parameter estimate with 500 replications in the error distributions of E_1 , E_2 and E_3 . In this time the dispersion parameter θ is known.

Error	(K,n)	method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
E_1	(10,10)	HRE	1.996(0.396)*	1.991(0.239)	0.990(0.245)	1.008(0.249)
		LSE	1.996(0.397)	1.989(0.237)	0.991(0.245)	1.008(0.251)
	(20,5)	HRE	1.980(0.329)	2.010(0.261)	1.002(0.247)	1.004(0.257)
		LSE	1.980(0.327)	2.010(0.261)	1.003(0.247)	1.004(0.259)
E_2	(10,10)	HRE	2.009(0.396)	2.016(0.251)	1.003(0.230)	1.004(0.232)
		LSE	2.008(0.392)	2.016(0.251)	1.004(0.230)	1.004(0.230)
	(20,5)	HRE	1.997(0.329)	1.997(0.257)	0.998(0.247)	1.009(0.253)
		LSE	1.997(0.330)	1.998(0.257)	0.998(0.247)	1.007(0.251)
E_3	(10,10)	HRE	2.031(0.483)	2.013(0.374)	0.967(0.387)	0.988(0.381)
		LSE	2.035(0.519)	2.012(0.417)	0.963(0.434)	0.987(0.437)
	(20,5)	HRE	2.014(0.454)	2.007(0.134)	0.996(0.399)	0.978(0.399)
		LSE	2.013(0.480)	2.011(0.459)	1.000(0.435)	0.971(0.436)

* Standard deviations are in parentheses.

Table 2: Results of parameter estimate with 500 replications in the error distributions of E_1 , E_2 and E_3 . In this time the dispersion parameter θ is unknown.

Error	(K, n_i)	method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}^2$
E_1	(10,10)	HRE	1.971(0.374)	1.990(0.251)	1.016(0.243)	1.014(0.221)	1.008
		LSE	1.971(0.373)	1.989(0.251)	1.016(0.241)	1.014(0.224)	1.127
	(20,5)	HRE	1.988(0.329)	1.994(0.261)	1.002(0.259)	1.002(0.245)	0.977
		LSE	1.989(0.327)	1.994(0.263)	1.002(0.255)	1.001(0.247)	1.125
	(20,10)	HRE	1.988(0.267)	2.008(0.161)	1.010(0.163)	0.992(0.167)	0.963
		LSE	1.988(0.269)	2.007(0.164)	1.010(0.164)	0.992(0.166)	1.088
E_2	(10,10)	HRE	1.975(0.383)	2.000(0.228)	0.996(0.241)	1.012(0.228)	1.000
		LSE	1.976(0.383)	1.999(0.228)	0.996(0.239)	1.011(0.226)	1.114
	(20,5)	HRE	1.985(0.345)	2.006(0.263)	0.987(0.253)	0.996(0.257)	0.979
		LSE	1.987(0.344)	2.006(0.265)	0.986(0.251)	0.996(0.257)	1.132
	(20,10)	HRE	2.011(0.269)	1.993(0.168)	0.996(0.169)	1.004(0.171)	0.990
		LSE	2.010(0.269)	1.993(0.166)	0.997(0.167)	1.004(0.171)	1.118
E_3	(10,10)	HRE	2.012(0.491)	2.006(0.375)	1.000(0.363)	0.991(0.397)	1.037
		LSE	2.006(0.519)	2.008(0.425)	1.003(0.409)	0.998(0.445)	0.404
	(20,5)	HRE	2.016(0.428)	1.972(0.392)	1.004(0.415)	0.984(0.418)	1.119
		LSE	2.032(0.449)	1.958(0.422)	0.989(0.443)	0.984(0.446)	0.430
	(20,10)	HRE	2.024(0.326)	1.988(0.264)	0.992(0.284)	0.999(0.258)	1.085
		LSE	2.029(0.354)	1.982(0.304)	0.990(0.323)	0.998(0.293)	0.453

Table 3: Comparisons of median prediction errors with HRE and LSE for cakes data and sleepstudy data

Data	method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	MPE	MPPE	MAPE	MSPE
cakes	HRE	2.251	0.006	0.003	4.7914	0.0185	0.2728	23.038
	LSE	2.247	0.006	1.142	5.0339	0.0209	0.2899	25.353
sleepstudy	HRE	5.532	0.033	0.012	31.1827	0.0077	0.1758	972.44
	LSE	5.532	0.033	0.899	31.3578	0.0085	0.1845	983.42